УДК [004.93+004.42]

Студ. В.А. Ворошень
Науч. рук. зав. кафедрой Д.М. Романенко
(кафедра информатики и веб-дизайна, БГТУ)

ОСОБЕННОСТИ ПРОГРАММНОЙ РЕАЛИЗАЦИИ КОНВЕРТАЦИИ АРХИВИРОВАННОГО XML С ЧАСТИЧНЫМ СОХРАНЕНИЕМ ФОРМАТИРОВАНИЯ

Существуют различные форматы данных, некоторые из которых при рассмотрении оказываются набором файлов формата .xml. К ним можно отнести форматы файлов, создаваемых офисными пакетами, например, .docx, .xlsx, .pptx, а также .odt, .ods и .odp.

Формат .docx является частью спецификации и стандарта для электронных документов OpenXML, разработанного корпорацией Microsoft. Данный формат представляет собой стандартизированный способ хранения документов: текстовых файлов, таблиц, изображений. Формат .odt является частью открытого набора стандартов ODF, который был разработан Sun Microsystems и стандартизован OASIS. Он используется в офисных пакетах, например, LibreOffice.

Оба формата расширены из XML.

Поставленной задачей было извлечение из документов форматов .docx и .odt текстового содержимого с сохранением такого форматирования, как выделение полужирным и курсивным начертанием, а также выравнивание текста. Для этого использовались средства языка PHP.

В случае обоих форматов алгоритм действий сходный:

1. получить сведения о файле;
2. разархивировать файл;
3. получить текстовое содержимое;
4. преобразовать текст в необходимый вид.

Получение метаданных файла необходимо, чтобы корректно работать с ним, учитывая уровни вложенности, имена папок и прочее.

Для разархивирования файлов удобно воспользоваться модулем ZipArchive. С его помощью возможно создать папку, где будет находится содержимое разархивированного файла, и извлечь его туда.

В зависимости от формата структура папок и файлов в созданной папке будет разниться. К счастью, весь текст с форматированием хранится в одном файле. В случае .doxc это файл document.xml, находящийся в папке word. В случае .odt – content.xml.

Работа с .docx состоит из нескольких этапов.

С помощью функции file_get_contents() в переменную записывается содержимое файла, представленное в виде XML-строки. Создаёт-

ся экземпляр класса SimpleXMLElement, который позволяет удобно взаимодействовать с такой строкой.

Важной особенностью XML является использование пространств имён, в которых содержатся элементы файла. В рассматриваемом формате это пространство имён «w».

С помощью метода children() в переменную записываются потомки данного пространства имён, в которых уже содержатся необходимые теги. Среди них тег «p», обозначающий абзац текста. Внутри него, в свою очередь, содержатся элементы «r» (run), напрямую содержащие текст (тег «t»). Необходимо извлечь все вхождения данного тега в файле, учитывая форматирование, которое обозначается тегами «b» и «i» для полужирного и курсивного начертаний соответственно. Сведения о выравнивании текста находятся в теге «pPr» в «jc». Благодаря доступу к дочерним элементам, возможно получить текстовые значения, чтобы в дальнейшем преобразовать их к нужному виду и, например, записать в базу данных. В отличие от действий для формата .docx, обработка .odt имеет нижеперечисленные особенности. Методы SimpleXMLElement могут использоваться и без создания экземпляра класса. Так с помощью функции simplexml_load_file() можно получить содержимое требуемого файла в виде XML-строки.

В отличие от формата .docx, .odt содержит несколько пространств имён на файл и необходимые элементы, описывающие содержимое и его оформление, содержатся в различных. Это менее удобно, поскольку приходится использовать функцию registerXPathNamespace(), регистрирующую в созданном объекте XML-строки необходимые пространства имён, которых может быть довольно много. Для извлечения текста с форматированием необходимо использовать три пространства имён, такие как «office», «styles» и «text». В «office» содержатся элементы верхнего уровня, тогда как стили, располагаемые отдельно, находятся в «styles», а текст – в «text».

Сперва регистрируется основное пространство имён, потом следующее за ним. Первыми извлекаются сведения о стилях текста, которые выгружаются в массив, содержащий их названия, а также значения форматирования. Текст же, в свою очередь, извлекается после стилей и приводится к нужному виду исходя из полученных данных.

Основные параметры форматирования в формате .odt заключаются в тег «span», благодаря чему их достаточно просто извлекать.

Таким образом, если возникает необходимость получения текста с сохранением его форматирования, необходимо учитывать особенности используемых форматов.